CLAIMS

1. Processing apparatus for generating classification data for text, the processing apparatus comprising:

identifying means for identifying semantic content bearing lexical units in data representing the text to be classified;

sequence determining means for determining sequences of the identified lexical units; and

classification data determining means for determining classification data as a score for the text to be classified with respect to each of a plurality of qualities by comparing the determined sequences of the identified lexical units with stored sequences of lexical units for training texts having scores associated therewith for a plurality of qualities.

2. Processing apparatus according to claim 1, wherein including storage means for storing the stored sequences of lexical units as at least one sequence of lexical units starting from each consecutive semantic content bearing lexical unit in data representing training text, and said sequence determining means is adapted to determine at least one sequence of lexical units starting from each consecutive semantic content bearing lexical unit in the text to be classified, and said classification data determining means is adapted to determine the scores by comparing said at least one sequence starting from each consecutive semantic content bearing lexical unit in data representing the text to be classified with said at least one stored sequence starting from each consecutive semantic content bearing lexical unit in data representing the training text.

3. Processing apparatus according to claim 2, wherein said at least one sequence of lexical units comprises a sequence of previous lexical units.

4. Processing apparatus according to claim 2, wherein said at least one sequence of lexical units comprise sequences of 1 to n lexical units, where n is an integer greater than 1.

5. Processing apparatus according to claim 2, wherein said sequence determining means is adapted to determine sequences of lexical units in which the first lexical unit in said at least one sequence is not a common lexical unit or a modifying lexical unit that modifies the meaning of a subsequent lexical unit, and subsequent lexical units in said at least one sequence can be a modifying lexical unit.

6. Processing apparatus according to claim 2, wherein said sequence determining means is adapted to determine said at least one sequence of lexical units starting at the beginning of each sentence in the text to be classified so that said at least one sequence of lexical units does not include lexical units from another sentence and sequences of lexical units starting with lexical units at the beginning of sentences can include identifiers in the sequence to identify that there is no word in a position in the sequence.

7. Processing apparatus according to claim 1, wherein said at least one sequence of lexical units further includes a single semantic content bearing lexical unit.

8. Processing apparatus according to claim 1, wherein said identifying means is adapted to identify semantic content bearing lexical units by rejecting common words, and to stem words to provide the semantic content bearing lexical units as word stems.

9. Processing apparatus according to claim 1, including storage means storing scores for training texts and sequence scores for sequences of lexical units indicating the occurrence of the sequences in the training texts, wherein said sequence determining means is adapted to determine sequence scores for sequences of lexical units in the text to be classified, and said classification data determining means is adapted to compare the sequence scores for the training text and for the text to be classified to determine the scores for the text to be classified.

10. Processing apparatus according to claim 9, wherein said storage means stores the sequence scores associated with scores for the training texts

11. Processing apparatus according to claim 10, wherein said storage means stores the sequence scores for groups of scores for the training texts, and said classification data determining means is adapted to determine a group score for each group by comparing the sequence scores for the training text and for the text to be classified, and to determine the scores for the text to be classified from the group scores

12. Processing apparatus according to claim 11, wherein the groups of scores comprise a mid range group of mid range scores and at least one other group of scores above and below the mid range group.

13. Processing apparatus according to claim 1, wherein said classification data determining means is adapted to determine the scores for the text to be classified by attaching more weight to the comparison of longer sequences of lexical units than to shorter sequences of lexical units.

14. A method of generating classification data for text, the method comprising:

identifying semantic content bearing lexical units in data representing the text to be classified;

determining sequences of the identified lexical units; and

determining means for determining classification data as a score for the text to be classified with respect to each of a plurality of qualities by comparing the determined sequences of the identified lexical units with stored sequences of lexical units for training texts having scores associated therewith for a plurality of qualities.

15. A method according to claim 14, wherein the stored sequences of lexical units are stored as at least one sequence of lexical units starting from each consecutive semantic content bearing lexical unit in data representing training text, at least one sequence of lexical units is determined starting from each consecutive semantic content bearing lexical unit in the text to be classified, and the scores are determined by comparing said at least one sequence starting from each consecutive semantic content bearing lexical unit in data representing the text to be classified with said at least one

stored sequence starting from each consecutive semantic content bearing lexical unit in data representing the training text.

16. A method according to claim 15, wherein said at least one sequence of lexical units comprises a sequence of previous lexical units.

17. A method according to claim 15, wherein said at least one sequence of lexical units comprise sequences of 1 to n lexical units, where n is an integer greater than 1.

18. A method according to claim 15, wherein sequences of lexical units are determined in which the first lexical unit in said at least one sequence is not a common lexical unit or a modifying lexical unit that modifies the meaning of a subsequent lexical unit, and subsequent lexical units in said at least one sequence can be a modifying lexical unit.

19. A method according to claim 15, wherein said at least one sequence of lexical units is determined starting at the beginning of each sentence in the text to be classified so that said at least one sequence of lexical units does not include lexical units from another sentence and sequences of lexical units starting with lexical units at the beginning of sentences can include identifiers in the sequence to identify that there is no word in a position in the sequence.

20. A method according to claim 14, wherein said at least one sequence of lexical units further includes a single semantic content bearing lexical unit.

21. A method according to claim 14, wherein semantic content bearing lexical units are identified by rejecting common words, and words are stemmed to provide the semantic content bearing lexical units as word stems.

22. A method according to claim 14, including storing scores for training texts and sequence scores for sequences of lexical units indicating the occurrence of the sequences in the training texts, wherein sequence scores for sequences of lexical units in

the text to be classified are determined, and the sequence scores for the training text are compared to the sequenced scores for the text to be classified to determine the scores for the text to be classified.

23. A method according to claim 22, wherein the sequence scores associated with scores for the training texts are stored.

24. A method according to claim 23, wherein the sequence scores for groups of scores for the training texts are stored, a group score is determined for each group by comparing the sequence scores for the training text and for the text to be classified, and the scores for the text to be classified are determined from the group scores

25. A method according to claim 24, wherein the groups of scores comprise a mid range group of mid range scores and at least one other group of scores above and below the mid range group.

26. A method according to claim 14, wherein the scores for the text to be classified are determined by attaching more weight to the comparison of longer sequences of lexical units than to shorter sequences of lexical units.

27. Processing apparatus for generating classification data for text, the processing apparatus comprising:

program memory storing processor readable program code; and

a processor for reading and executing the program code;

wherein the program code comprises code to control the processor to:

identify semantic content bearing lexical units in data representing the text to be classified;

determine sequences of the identified lexical units; and

determine means for determining classification data as a score for the text to be classified with respect to each of a plurality of qualities by comparing the determined sequences of the identified lexical units with stored sequences of lexical units for training texts having scores associated therewith for a plurality of qualities.

28. Processing apparatus according to claim 27, including storage storing the stored sequences of lexical units as at least one sequence of lexical units starting from each consecutive semantic content bearing lexical unit in data representing training text, wherein the program code comprises code to control the processor to determine at least one sequence of lexical units starting from each consecutive semantic content bearing lexical unit in the text to be classified, and to determine the scores by comparing said at least one sequence starting from each consecutive semantic content bearing lexical unit in data representing the text to be classified with said at least one stored sequence starting from each consecutive semantic content bearing lexical unit in data representing the training text.

29. Processing apparatus according to claim 28, wherein said at least one sequence of lexical units comprises a sequence of previous lexical units.

30. Processing apparatus according to claim 28, wherein said at least one sequence of lexical units comprise sequences of 1 to n lexical units, where n is an integer greater than 1.

31. Processing apparatus according to claim 28, wherein the program code comprises code to control the processor to determine sequences of lexical units in which the first lexical unit in said at least one sequence is not a common lexical unit or a modifying lexical unit that modifies the meaning of a subsequent lexical unit, and subsequent lexical units in said at least one sequence can be a modifying lexical unit.

32. Processing apparatus according to claim 28, wherein the program code comprises code to control the processor to determine said at least one sequence of lexical units starting at the beginning of each sentence in the text to be classified so that said at least one sequence of lexical units does not include lexical units from another sentence and sequences of lexical units starting with lexical units at the beginning of sentences can include identifiers in the sequence to identify that there is no word in a position in the sequence.

33. Processing apparatus according to claim 27, wherein said at least one sequence of lexical units further includes a single semantic content bearing lexical unit.

34. Processing apparatus according to claim 27, wherein the program code comprises code to control the processor to identify semantic content bearing lexical units by rejecting common words, and words are stemmed to provide the semantic content bearing lexical units as word stems.

35. Processing apparatus according to claim 27, wherein the program code comprises code to control the processor to store scores for training texts and sequence scores for sequences of lexical units indicating the occurrence of the sequences in the training texts, to determine sequence scores for sequences of lexical units in the text to be classified, and to compare the sequence scores for the training text to the sequenced scores for the text to be classified to determine the scores for the text to be classified.

36. Processing apparatus according to claim 35, wherein the sequence scores associated with scores for the training texts are stored.

37. Processing apparatus according to claim 36, wherein the program code comprises code to control the processor to store the sequence scores for groups of scores for the training texts, to determine a group score for each group by comparing the sequence scores for the training text and for the text to be classified, and to determine the scores for the text to be classified from the group scores

38. Processing apparatus according to claim 37, wherein the groups of scores comprise a mid range group of mid range scores and at least one other group of scores above and below the mid range group.

39. Processing apparatus according to claim 27, wherein the program code comprises code to control the processor to determine the scores for the text to be

classified by attaching more weight to the comparison of longer sequences of lexical units than to shorter sequences of lexical units.

40. Processing apparatus for generating classification data for text, the processing apparatus comprising:

identifying means for identifying semantic content bearing lexical units in data representing the text to be classified; and

classification data determining means for determining classification data as a score for the text to be classified with respect to each of a plurality of qualities by comparing the identified lexical units with stored lexical units having a distribution of lexical scores associated therewith for each of a plurality of qualities.

41. Processing apparatus according to claim 40, including storage means storing said distribution of lexical scores for each of the plurality of qualities, the distribution having been obtained from training data.

42. Processing apparatus according to claim 40, wherein said classification data determining means is adapted to determine the score for the text to be classified by statistical analysis of the result of the comparison.

43. Processing apparatus according to claim 40, including sequence determining means for determining sequences of the identified lexical units; wherein said classification data determining means is adapted to determine the score by comparing the determined sequences of the identified lexical units with stored sequences of lexical units for training texts having score distributions associated therewith for the plurality of qualities.

44. Processing apparatus for generating classification data for text, the processing apparatus comprising:

program memory storing processor readable program code; and

a processor for reading and executing the program code;

wherein the program code comprises code to control the processor to:

identify semantic content bearing lexical units in data representing the text to be classified; and

determine classification data as a score for the text to be classified with respect to each of a plurality of qualities by comparing the identified lexical units with stored lexical units having a distribution of lexical scores associated therewith for each of a plurality of qualities.

45. Processing apparatus according to claim 44, including storage storing said distribution of lexical scores for each of the plurality of qualities, the distribution having been obtained from training data.

46. Processing apparatus according to claim 44, wherein the program code comprises code to control the processor to determine the score for the text to be classified by statistical analysis of the result of the comparison.

47. Processing apparatus according to claim 44, wherein the program code comprises code to control the processor to determine sequences of the identified lexical units; and to determine the score by comparing the determined sequences of the identified lexical units with stored sequences of lexical units for training texts having score distributions associated therewith for the plurality of qualities.

48. A method of generating classification data for text, the method comprising:

identifying semantic content bearing lexical units in data representing the text to be classified; and

determining classification data as a score for the text to be classified with respect to each of a plurality of qualities by comparing the identified lexical units with stored lexical units having a distribution of lexical scores associated therewith for each of a plurality of qualities.

49. A method according to claim 48, including storing said distribution of lexical scores for each of the plurality of qualities, the distribution having been obtained from training data.

50. A method according to claim 48, wherein the score for the text to be classified is determined by statistical analysis of the result of the comparison.

51. A method according to claim 48, including determining sequences of the identified lexical units; wherein the score is determined by comparing the determined sequences of the identified lexical units with stored sequences of lexical units for training texts having score distributions associated therewith for the plurality of qualities.

52. Processing apparatus for generating classification data in a hierarchical structure for text, the processing apparatus comprising:

the processing apparatus according to claim 1 or claim 40;

wherein said classification data determining means is adapted select a quality having the highest score and to repeat the determination of a score for a set of qualities dependent upon the selected quality.

53. Processing apparatus according to claim 52, wherein said classification data determining means is adapted to use a sub set of the stored training texts dependant upon the selected quality for the repeated determination.

54. A method of generating classification data in a hierarchical structure for text, the method comprising:

the method of claim 14 or claim 48; and

selecting a quality having the highest score and repeating the determination of a score for a set of qualities dependent upon the selected quality.

55. A method according to claim 54, wherein a sub set of the stored training texts dependant upon the selected quality is used for the repeated determination

56. Processing apparatus according to claim 1 or claim 40, including training data modifying means for modifying the training data using the classification data if confidence in the classification is high.

57. A method of claim 14 or claim 48, including modifying the training data using the classification data if confidence in the classification is high.

58. An automatic text classification system comprising:

means for extracting word stems and word stem sequences from data representing a text to be classified;

means for calculating a probability value for the text to be classified with respect to each of a plurality of qualities based on a correlation between (i) the extracted word stems and word stem sequences and (ii) predetermined training data.

59. The automatic text classification system according to claim 58, wherein each quality is represented by an axis whose two end points correspond to mutually exclusive characteristics.

60. The automatic text classification system according to claim 59, wherein the probability value with respect to each of the plurality of qualities is converted into a score on each axis indicating a likelihood of the text having one or the other of the mutually exclusive characteristics.

61. The automatic text classification system according to claim 58, wherein the training data is derived from a plurality of training texts that have been pre-classified with respect to each of the plurality of qualities, and the training data comprises a distribution value of each word stem and each word stem sequence in each of the plurality of training texts with respect to each of the plurality of qualities.

62. The automatic text classification system according to claim 61, wherein:

each quality is represented by an axis that is divided into a plurality of groups and whose two end points correspond to mutually exclusive characteristics;

each of the training texts has been pre-classified into one of the groups on each axis;

the training data comprises a database of, for each group on each axis, the distribution value of each word stem and word stem sequence in each training text with respect to the one group on each axis into which each training text has been pre-classified;

the distribution values represent a probability of each word stem and word stem sequence existing in a group on a given axis; and

the probability values of the text to be classified represent a probability of the text being classified in each group on each axis.

63. The automatic text classification system according to claim 62, wherein each of the training texts has been pre-classified with a specific score on each axis, and each group on each axis comprises a predetermined range of scores.

64. The automatic text classification system according to claim 63, wherein the training texts are selected so that the pre-classified scores are distributed along each axis between a Bell curve and a flat distribution.

65. The automatic text classification system according to claim 63, wherein:

each axis is divided into a first group, a neutral second group, and a third group; and

the neutral second group with respect to the pre-classification of the training texts is broader than the neutral second group with respect to the text to be classified, so that the probability values of the text to be classified are more likely to be converted into scores which fall on an appropriate side of each axis.

66. The automatic text classification system according to claim 58, wherein:

each word stem is a main stem word that is not a common word;

a modifying word is a common word that adds meaning to a main stem word; and

each word stem sequence comprises a main stem word and one or more previous words that are either modifying words or other main stem words.

67. The automatic text classification system according to claim 66, wherein the probability values are calculated such that a correlation between an extracted triple word stem sequence with the training data is more heavily weighted than a correlation between an extracted double word stem sequence with the training data, and such that a correlation between an extracted double word stem sequence with the training data is more heavily weighted than a correlation between a single extracted word stem with the training data.

68. A system for producing training data comprising:

means for extracting word stems and word stem sequences from each of a plurality of training texts that have been pre-classified with respect to each of a plurality of qualities; and

means for calculating a distribution value of each extracted word stem and word stem sequence in each training text with respect to each of the plurality of qualities.

69. The system for producing training data according to claim 68, wherein each quality is represented by an axis whose two end points correspond to mutually exclusive characteristics.

70. The system for producing training data according to claim 68, wherein:

each quality is represented by an axis that is divided into a plurality of groups and whose two end points correspond to mutually exclusive characteristics;

each of the training texts has been pre-classified into one of the groups on each axis;

the training data comprises a database of, for each group on each axis, a distribution value of each word stem and word stem sequence in each training text with respect to the one group on each axis into which each training text has been pre-classified; and

the distribution values represent a probability of each word stem and word stem sequence existing in a given group on a given axis.

71. The system for producing training data according to claim 70, wherein each of the training texts has been pre-classified with a specific score on each axis, and each group on each axis comprises a predetermined range of scores.

72. The system for producing training data according to claim 71, wherein the training texts are selected so that the pre-classified scores are distributed along each axis between a Bell curve and a flat distribution.

73. The system for producing training data according to claim 68, wherein:

each word stem is a main stem word that is not a common word;

a modifying word is a common word that adds meaning to a main stem word; and

each word stem sequence comprises a main stem word and one or more previous words that are either modifying words or other main stem words.

74. The system for producing training data according to claim 68, further comprising:

means for, after a plurality of new texts have been classified with respect to the plurality of qualities using the training data, selecting a number of the new texts that have been classified with a predetermined degree of probability with respect to at least one of the plurality of qualities;

means for extracting word stems and word stem sequences from each of the selected new texts; and

means for one of (i) recalculating the distribution value of each extracted word stem and word stem sequence which is already present in the training data, and (ii) calculating an initial distribution value of each extracted word stem and word stem sequence which is not already present in the training data.

75. A retrieval system comprising:

means for accessing a data store comprising a plurality of word stems and word stem sequences that have been extracted from a plurality of texts, a plurality of identifiers associating each word stem and word stem sequence with at least one of the plurality of texts, and correlation data between (i) each word stem and word stem sequence and (ii) each of a plurality of qualities in terms of which the plurality of texts have been classified;

means for receiving user preference data in terms of at least one of the plurality of qualities;

means for identifying word stems and word stem sequences corresponding to the user preference data based on the correlation data stored in the data store using fuzzy logic; and

means for identifying at least one of the plurality of texts that best matches the user preference data based on the identified word stems and word stem sequences and the plurality of identifiers stored in the data store.

76. The retrieval system according to claim 75, wherein each quality is represented by an axis whose two end points represent mutually exclusive characteristics.

77. The retrieval system according to claim 75, wherein:

each quality is represented by an axis that is divided into a plurality of groups and whose two end points correspond to mutually exclusive characteristics;

each of the plurality of texts has been classified into one of the groups on each axis;

the correlation data comprises, for each group on each axis, a distribution value of each word stem and word stem sequence in each text with respect to the one group on each axis into which each text has been classified; and

the distribution values represent a probability of each word stem and word stem sequence existing in a given group on a given axis.

78. The retrieval system according to claim 77, wherein:

each word stem is a main stem word that is not a common word;

a modifying word is a common word that adds meaning to a main stem word; and

each word stem sequence comprises a main stem word and one or more previous words that are either modifying words or other main stem words.

79. The retrieval system according to claim 75, further comprising a graphical user interface for enabling input of the user preference data.

80. A system for producing training data comprising:

means for identifying lexical units and lexical unit sequences from each of a plurality of training texts that have been pre-classified with respect to each of a plurality of qualities; and

means for calculating a distribution value of each identified lexical unit and lexical unit sequence in each training text with respect to each of the plurality of qualities.

81. A method of producing training data comprising:

identifying lexical units and lexical unit sequences from each of a plurality of training texts that have been pre-classified with respect to each of a plurality of qualities; and

calculating a distribution value of each identified lexical unit and lexical unit sequence in each training text with respect to each of the plurality of qualities

82. A carrier medium carrying computer readable code for controlling a processor to carry out the method of any one of claims 14 to 26, 48 to 51, 54, 55 or 57.

83. A carrier medium carrying computer readable code for controlling a computer to function as the system as claimed in any one of the claims 58 to 79.